Random

Shahirah Idris

2023-03-17

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Setting up my environment

Notes: Setting up my R environment by loading the 'tidyverse', 'dyplr', 'janitor', 'ggplot2' and 'skimr' pacakges.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.0     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.1     ✓ tibble    3.1.8
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the  ]8;;http://conflicted.r-lib.org/ conflicted package ]8;;  to force all conflict
s to become errors
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(ggplot2)
library(skimr)
library(dplyr)
```

Import and read the flavors_of_cocoa.csv file

Notes: Import and read the csv file downloaded form Kaggle and save it as a dataframe. (chocolate_df)

```
chocolate_df <- read_csv("flavors_of_cacao.csv")
```

```
## Rows: 1795 Columns: 9
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (6): Company
## (Maker-if known), Specific Bean Origin
## or Bar Name, Cocoa
## ...
## dbl (3): REF, Review
## Date, Rating
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Show the first few rows of the dataset.

Notes: There are about 9 columns.

```
head(chocolate_df)
```

```
## # A tibble: 6 × 9
##   Company \n(Make…¹ Speci…²   REF Revie…³ Cocoa…⁴ Compa…⁵ Rating Bean\…⁶ Broad…⁷
##   <chr>             <chr>   <dbl>   <dbl> <chr>   <chr>    <dbl> <chr>   <chr>
## 1 A. Morin          Agua G…  1876    2016 63%     France    3.75         Sao To…
## 2 A. Morin          Kpime    1676    2015 70%     France    2.75         Togo
## 3 A. Morin          Atsane   1676    2015 70%     France    3            Togo
## 4 A. Morin          Akata    1680    2015 70%     France    3.5          Togo
## 5 A. Morin          Quilla   1704    2015 70%     France    3.5          Peru
## 6 A. Morin          Carene…  1315    2014 70%     France    2.75 Criollo Venezu…
## # … with abbreviated variable names ¹`Company \n(Maker-if known)`,
## #   ²`Specific Bean Origin\nor Bar Name`, ³`Review\nDate`, ⁴`Cocoa\nPercent`,
## #   ⁵`Company\nLocation`, ⁶`Bean\nType`, ⁷`Broad Bean\nOrigin`
```

Show the column names.

Notes: To see if the names of the column are easily readable.

```
colnames(chocolate_df)
```

```
## [1] "Company \n(Maker-if known)"        "Specific Bean Origin\nor Bar Name"
## [3] "REF"                               "Review\nDate"
## [5] "Cocoa\nPercent"                    "Company\nLocation"
## [7] "Rating"                            "Bean\nType"
## [9] "Broad Bean\nOrigin"
```

Using spec() to extract the full column specification

from a tibble created by readr.

```
spec(chocolate_df)
```

```
## cols(
##   `Company
## (Maker-if known)` = col_character(),
##   `Specific Bean Origin
## or Bar Name` = col_character(),
##   REF = col_double(),
##   `Review
## Date` = col_double(),
##   `Cocoa
## Percent` = col_character(),
##   `Company
## Location` = col_character(),
##   Rating = col_double(),
##   `Bean
## Type` = col_character(),
##   `Broad Bean
## Origin` = col_character()
## )
```

Clean the columns' names with clean_names().

Notes: Cleaning the column names will make the analysis more accessible.

```
unclean_flavors_df <-
  chocolate_df %>%
  clean_names()
```

Display the column names of new dataframe.

Notes: Displaying the column names with col_names() function after cleaning them for improved readability.

```
colnames(unclean_flavors_df)
```

```
## [1] "company_maker_if_known"          "specific_bean_origin_or_bar_name"
## [3] "ref"                             "review_date"
## [5] "cocoa_percent"                   "company_location"
## [7] "rating"                          "bean_type"
## [9] "broad_bean_origin"
```

Rename the column name

Notes: Renaming the column names for consistency and clarity.

```
flavors_df <-
  unclean_flavors_df %>%
  rename(company=company_maker_if_known)
```

View the dataframe

```
head(flavors_df)
```

```
## # A tibble: 6 × 9
##   company  specific_bean_…¹   ref revie…² cocoa…³ compa…⁴ rating bean_…⁵ broad…⁶
##   <chr>    <chr>            <dbl>   <dbl> <chr>   <chr>    <dbl> <chr>   <chr>
## 1 A. Morin Agua Grande       1876    2016 63%     France    3.75         Sao To…
## 2 A. Morin Kpime             1676    2015 70%     France    2.75         Togo
## 3 A. Morin Atsane            1676    2015 70%     France    3            Togo
## 4 A. Morin Akata             1680    2015 70%     France    3.5          Togo
## 5 A. Morin Quilla            1704    2015 70%     France    3.5          Peru
## 6 A. Morin Carenero          1315    2014 70%     France    2.75 Criollo Venezu…
## # … with abbreviated variable names ¹specific_bean_origin_or_bar_name,
## #   ²review_date, ³cocoa_percent, ⁴company_location, ⁵bean_type,
## #   ⁶broad_bean_origin
```

Analyzing the dataframe for insights.

Notes: To collect and analyze data on the latest chocolate ratings and information on which countries produce the highest-rated bars of super dark chocolate.

```
trimmed_flavors_df <-
  flavors_df %>%
  select(rating, cocoa_percent, company_location)
```

View the first few rows of the dataframe.

```
head(trimmed_flavors_df)
```

```
## # A tibble: 6 × 3
##   rating cocoa_percent company_location
##    <dbl> <chr>         <chr>
## 1   3.75 63%           France
## 2   2.75 70%           France
## 3   3    70%           France
## 4   3.5  70%           France
## 5   3.5  70%           France
## 6   2.75 70%           France
```

Summary on the rating of the chocolate.

Notes: Using summarize() and mean() functions.

```
trimmed_flavors_df %>% summarize(mean(rating))
```

```
## # A tibble: 1 × 1
##   `mean(rating)`
##            <dbl>
## 1           3.19
```

Filtering the dataframe to only show data of the chocolate with high cocoa percent and high rating.

Notes: Chocolate with at least 75% cocoa percent is super dark chocolate. High rating chocolate has at least 3.9 points.

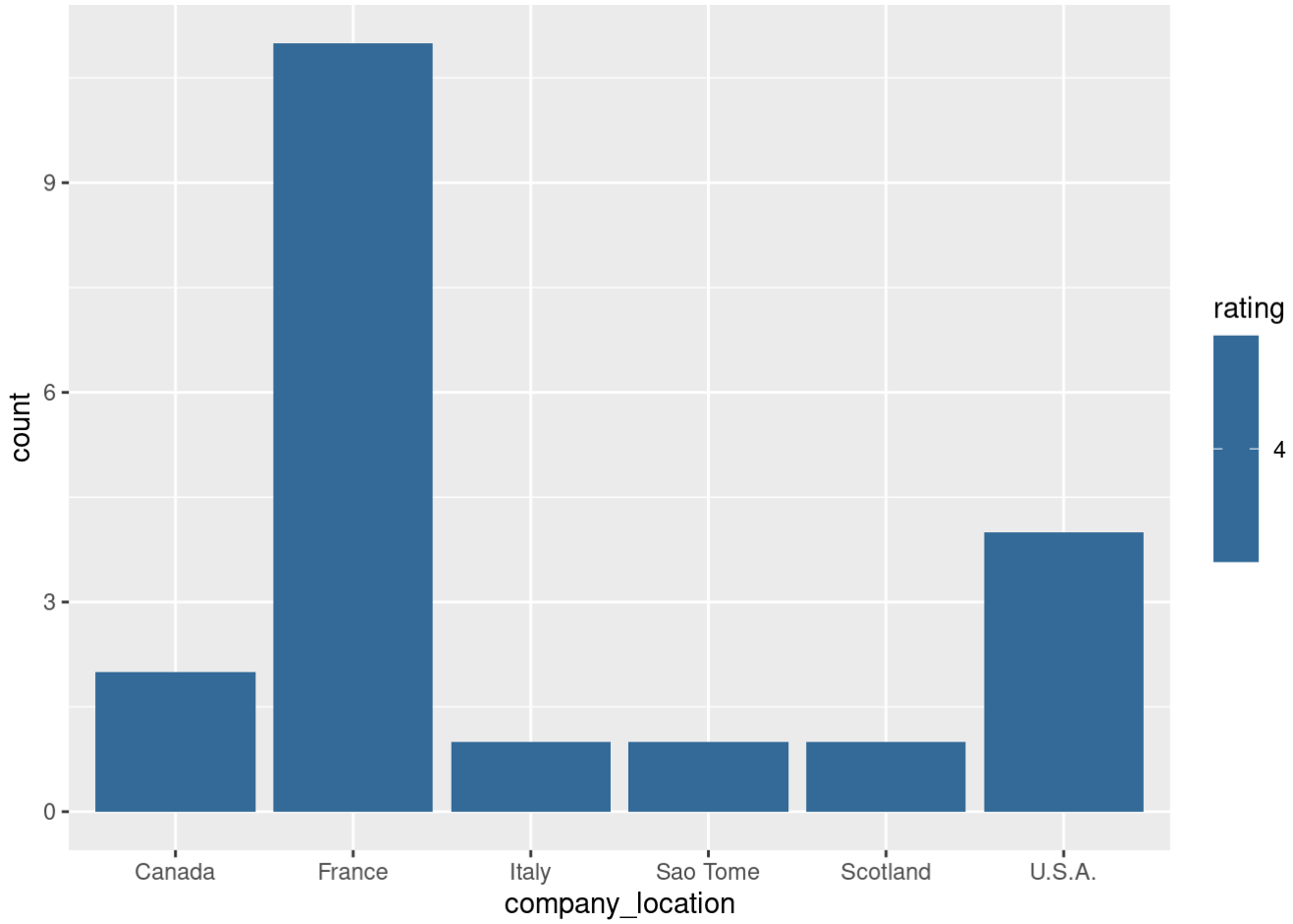
```
best_trimmed_flavors_df <- trimmed_flavors_df %>%
  filter(cocoa_percent >= "75", rating >= "3.9")

head(best_trimmed_flavors_df)
```

```
## # A tibble: 6 × 3
##   rating cocoa_percent company_location
##    <dbl> <chr>         <chr>
## 1      4 75%           Italy
## 2      4 75%           France
## 3      4 75%           France
## 4      4 75%           France
## 5      4 75%           France
## 6      4 75%           France
```

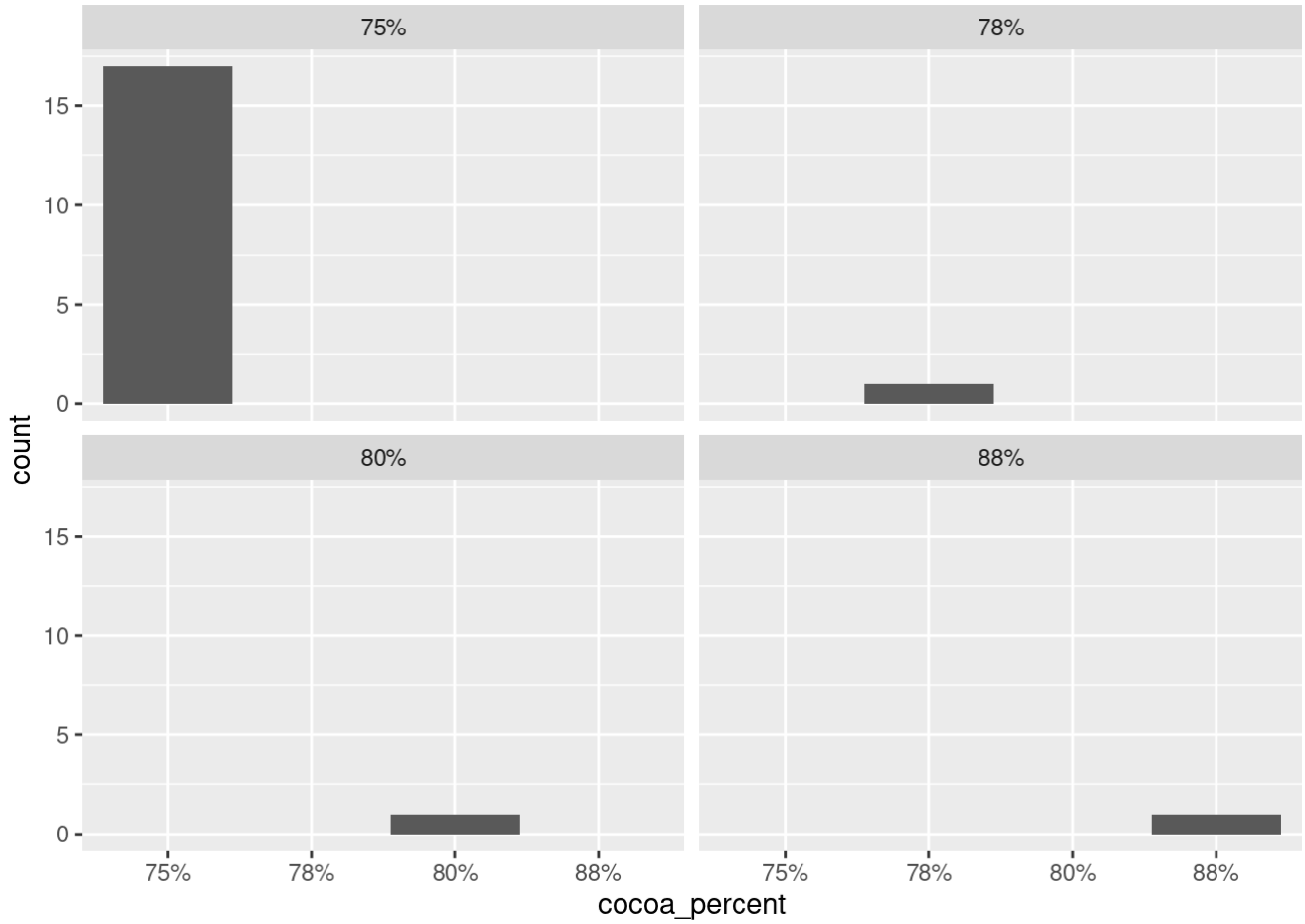
Data Visualization to show which countries produce the highest-rated bars of super dark chocolate.

```
ggplot(data = best_trimmed_flavors_df) +
  geom_bar(mapping = aes(x = company_location, fill = rating))
```

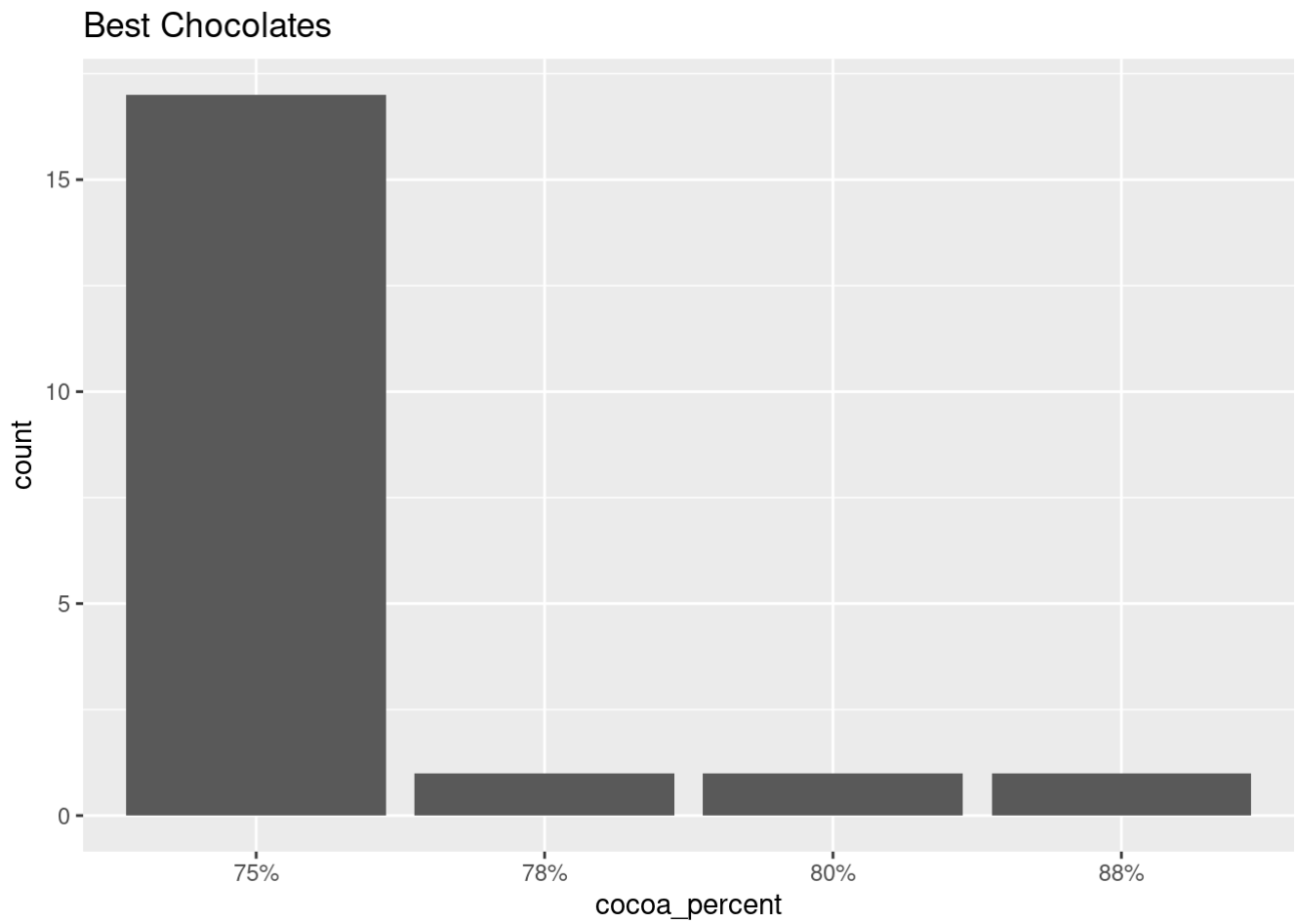


From the visualization above, France and U.S.A produce the highest rated chocolate bars.

```
ggplot(data = best_trimmed_flavors_df) +
  geom_bar(mapping = aes(x = cocoa_percent)) +
  facet_wrap(~cocoa_percent)
```

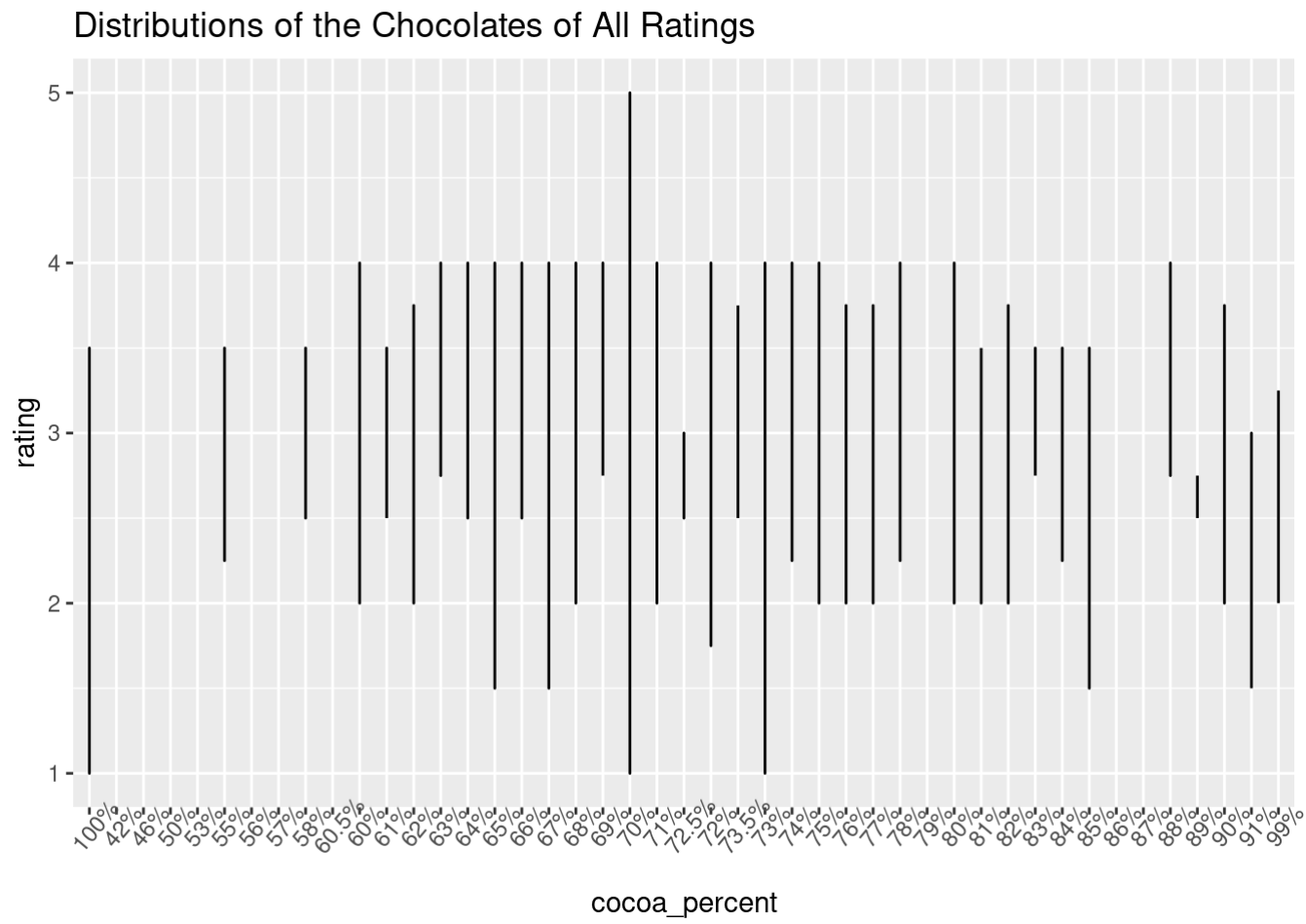



```
ggplot(data = best_trimmed_flavors_df) +
  geom_bar(mapping = aes(x = cocoa_percent)) +
  labs(title = "Best Chocolates")
```



Chocolate bars having 75% cocoa percent is the most produced chocolates.

```
ggplot(data = trimmed_flavors_df) +
  geom_line(mapping = aes(x = cocoa_percent, y = rating)) +
  theme(axis.text.x = element_text(angle = 50)) +
  labs(title = "Distributions of the Chocolates of All Ratings")
```



Showing the min, max and average rating of the chocolate bars produced by respective countries.

```
chocolate_summary <-
  trimmed_flavors_df %>%
  group_by(company_location) %>%
  summarise(average_chocolate_rating = mean(rating),
            min_rating = min(rating),
            max_rating = max(rating))
```

Find the country with highest and lowest average rating of chocolate bars.

Notes: The min and max average rating for the chocolate bars are 2.5 and 3.75 respectively.

```
chocolate_summary %>%
  summarise(min_average_rating = min(average_chocolate_rating),
            max_average_rating = max(average_chocolate_rating))
```

```
## # A tibble: 1 × 2
##   min_average_rating max_average_rating
##                <dbl>              <dbl>
## 1                2.5               3.75
```

```
chocolate_summary %>%
  filter(average_chocolate_rating == 2.5)
```

```
## # A tibble: 1 × 4
##   company_location average_chocolate_rating min_rating max_rating
##   <chr>                                <dbl>      <dbl>      <dbl>
## 1 India                                  2.5        2.5        2.5
```

```
chocolate_summary %>%
  filter(average_chocolate_rating == 3.75)
```

```
## # A tibble: 1 × 4
##   company_location average_chocolate_rating min_rating max_rating
##   <chr>                                <dbl>      <dbl>      <dbl>
## 1 Chile                                 3.75       3.75       3.75
```

Chile has the highest average chocolate rating at 3.75 and India has the lowest chcocolate ratio at 2.5.